TIME SERIES FORECASTING –ROSE WINE

DSBA

SUBMITTED BY SAUMYA JAIN (PGP-DSBA)

Contents

List of Tables:

List of Plots:

Problem:

For this assignment, the data of different types of wine sales in the 20th century is to be analysed. Both data are from the same company but of different wines. As an analyst in the ABC Estate Wines, you are tasked to analyse and forecast Wine Sales in the 20th century.

Data set for the Problem: [Rose.csv]

1. Read the data as an appropriate Time Series data and plot the data.

- The dataset given to us that contains the information of sales of rose wine. The excel has 187 rows and 1 column.
- We have also set index to be Year Month.
-
- The description of the dataset is as below.

	count	mean	std	min	25%	50%	75%	max
Rose	185.0	90.394595	39.175344	28.0	63.0	86.0	112.0	267.0

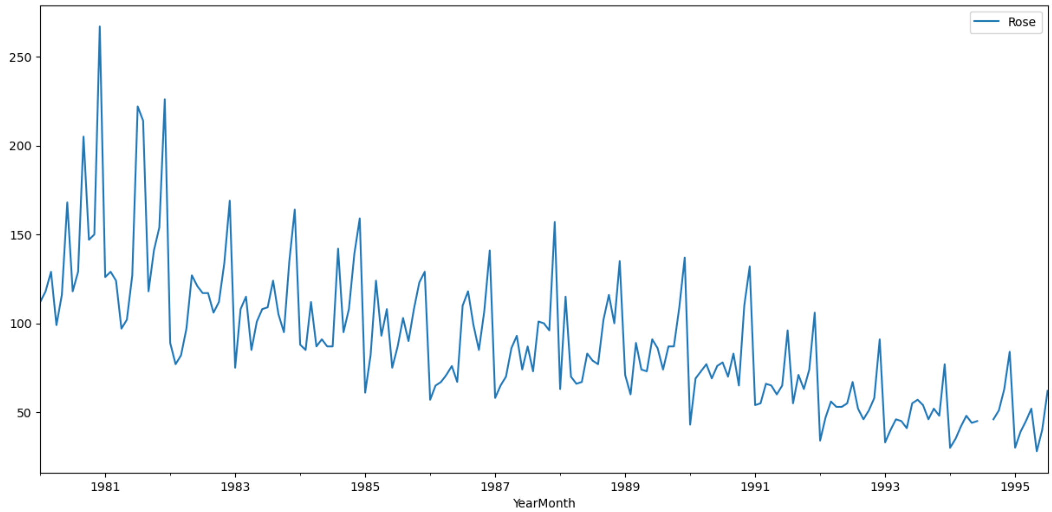
- We have seen the top 5 rows of the dataset shared as shown below:

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0

- We have seen the last 5 rows of the dataset shared as shown below:

	Rose
YearMonth	
1995-03-01	45.0
1995-04-01	52.0
1995-05-01	28.0
1995-06-01	40.0
1995-07-01	62.0

- Plot the graph:



2. Perform appropriate Exploratory Data Analysis to understand the data and perform decomposition.

- Datatypes of the data present in Rose wine dataset.

```
YearMonth     object
Rose         float64
dtype: object
```

- Check for null values present in the given dataset and found that there are 2 null values present in dataset as shown below:

```
Rose    2
dtype: int64
```

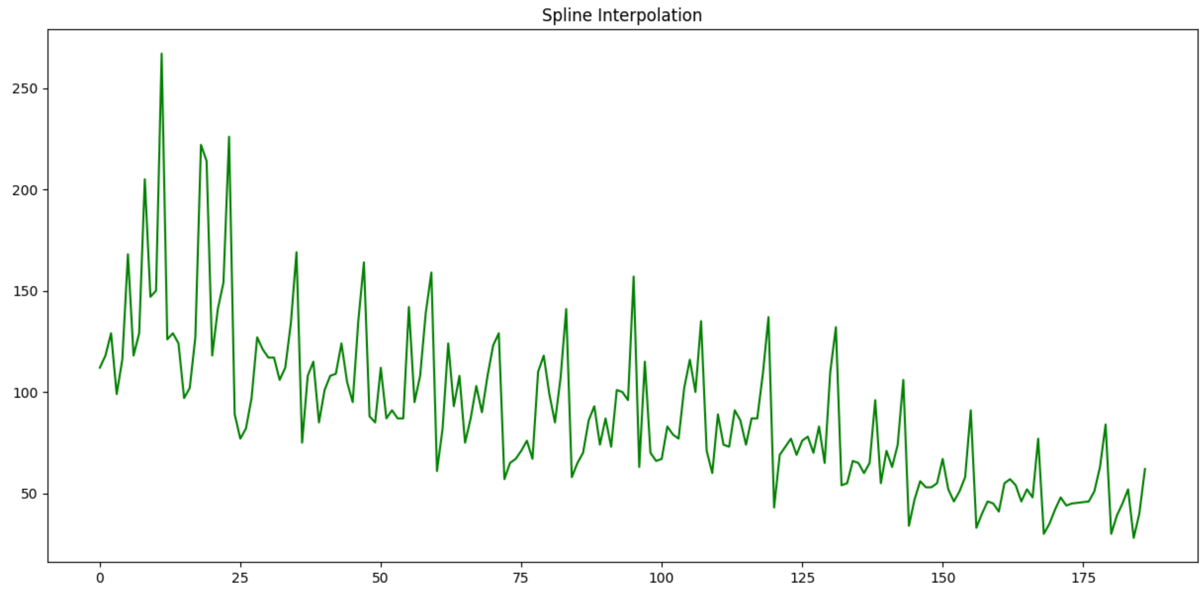
- The null values that is present in dataset is as below

YearMonth	Rose
1994-07-01	NaN
1994-08-01	NaN

- The null values need to be imputed as in Time series data analysis we can’t proceed with forecasting with the dataset with null values as it will affect the forecasting.
- For the same we need to use SPLINE for null values interpolation as SPLINE also work on large set and complex data present as it is more accurate. Interpolation technique.
- After using SPLINE technique there is no null values present in dataset as shown below.

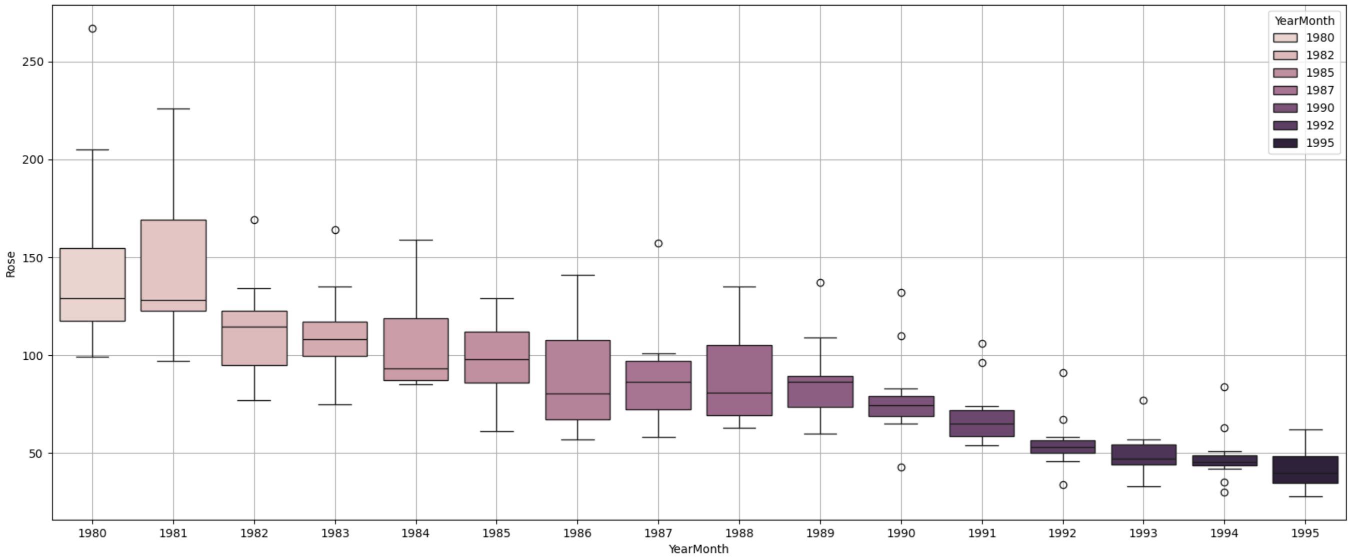
```
Rose    0
dtype: int64
```

- Graph after SPLINE Interpolation



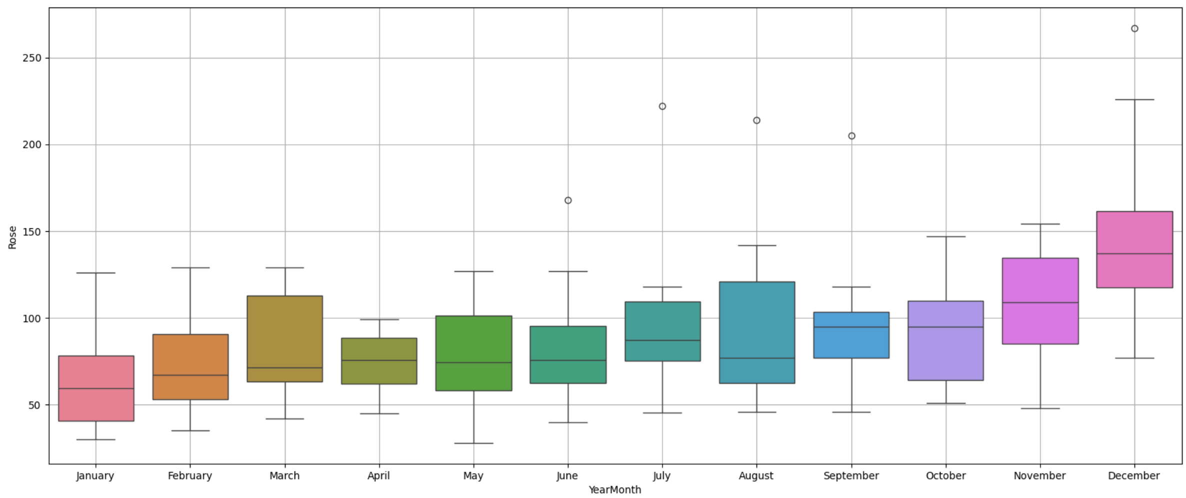
- **Boxplot Yearly:**

1. The boxplot yearly shows that the there is peak in 1980-1981 . Though outliers are also present in mostly all years.



- **Monthly Boxplot**

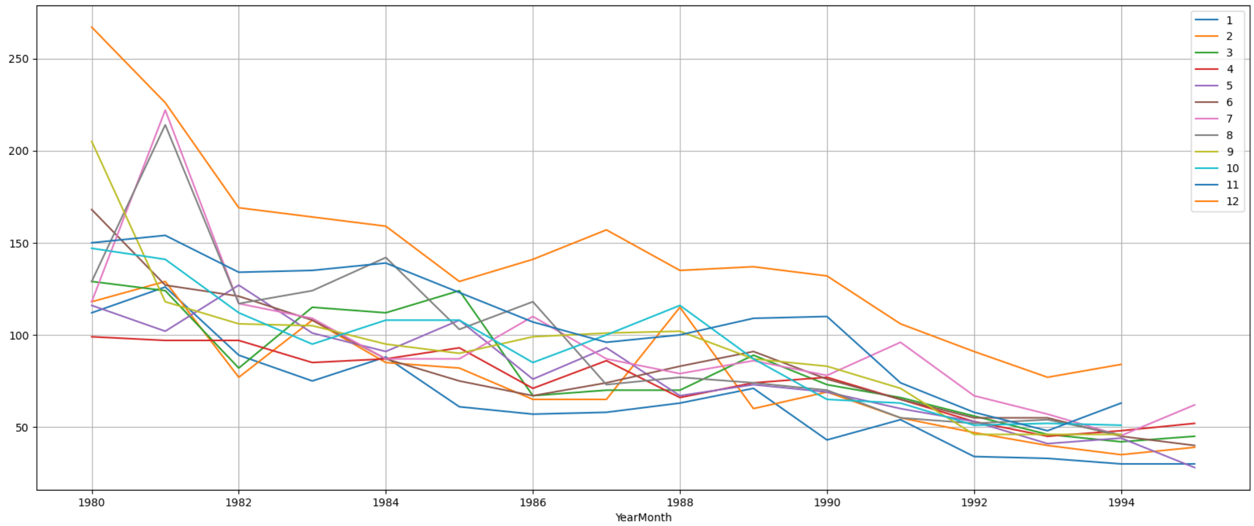
1. From the graph we can inference that the sales of rose wine is mostly high in December and lowest in January. Outliers are present in June, July, August , September and December.

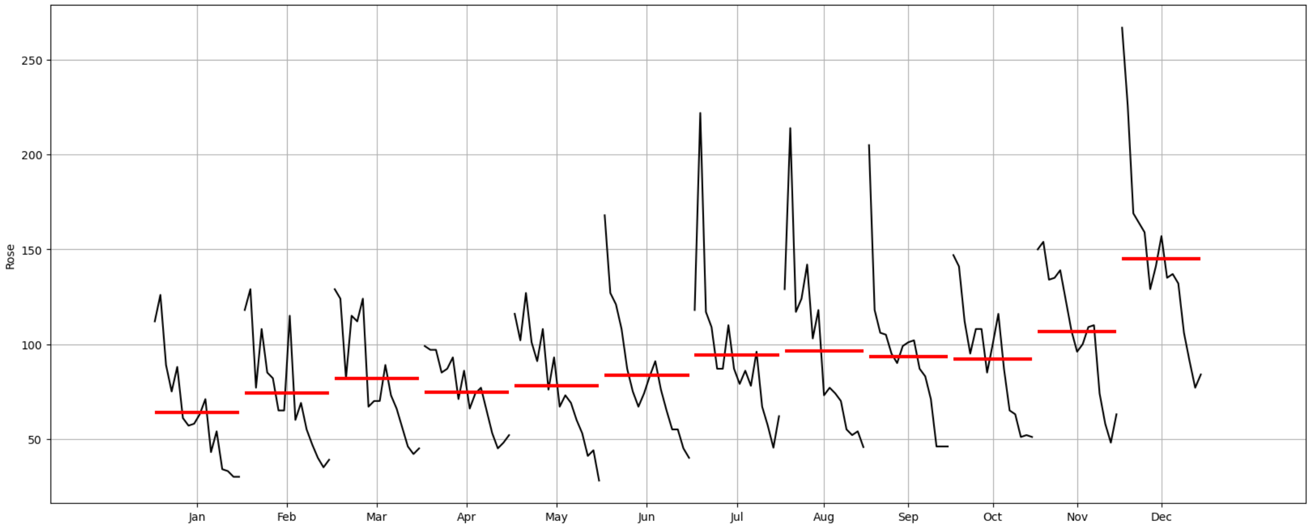


- ## Graph for Monthly Sales over the years:

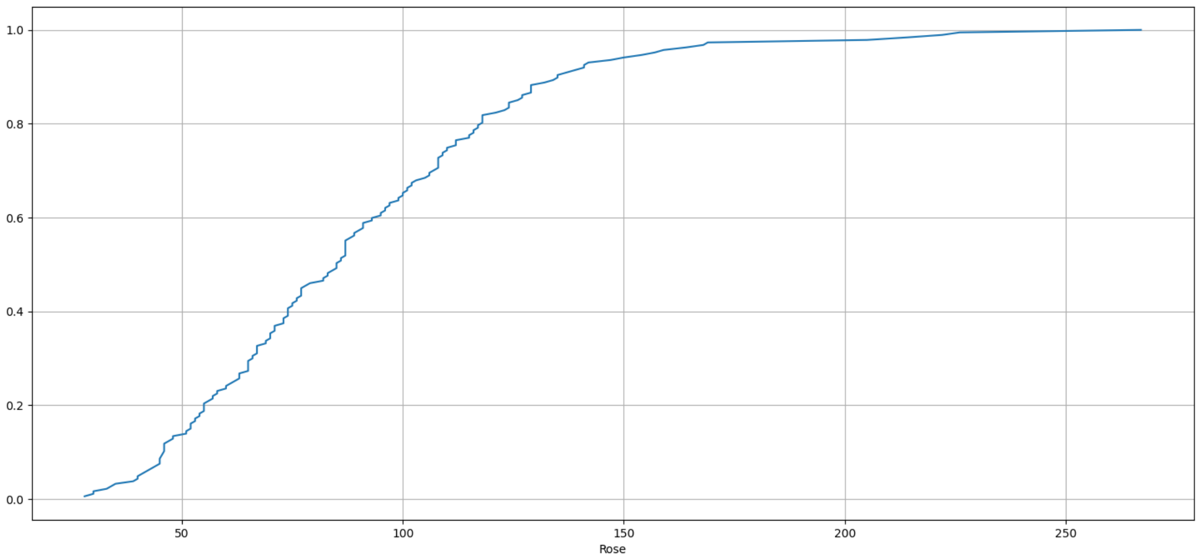
The sales of Rose wine is highest in 12th Month i.e. December and the year 1981 was the year with the highest number of sales.

YearMonth	1	2	3	4	5	6	7	8	9	10	11	12
YearMonth												
1980	112.0	118.0	129.0	99.0	116.0	168.0	118.000000	129.000000	205.0	147.0	150.0	267.0
1981	126.0	129.0	124.0	97.0	102.0	127.0	222.000000	214.000000	118.0	141.0	154.0	226.0
1982	89.0	77.0	82.0	97.0	127.0	121.0	117.000000	117.000000	106.0	112.0	134.0	169.0
1983	75.0	108.0	115.0	85.0	101.0	108.0	109.000000	124.000000	105.0	95.0	135.0	164.0
1984	88.0	85.0	112.0	87.0	91.0	87.0	87.000000	142.000000	95.0	108.0	139.0	159.0
1985	61.0	82.0	124.0	93.0	108.0	75.0	87.000000	103.000000	90.0	108.0	123.0	129.0
1986	57.0	65.0	67.0	71.0	76.0	67.0	110.000000	118.000000	99.0	85.0	107.0	141.0
1987	58.0	65.0	70.0	86.0	93.0	74.0	87.000000	73.000000	101.0	100.0	96.0	157.0
1988	63.0	115.0	70.0	66.0	67.0	83.0	79.000000	77.000000	102.0	116.0	100.0	135.0
1989	71.0	60.0	89.0	74.0	73.0	91.0	86.000000	74.000000	87.0	87.0	109.0	137.0
1990	43.0	69.0	73.0	77.0	69.0	76.0	78.000000	70.000000	83.0	65.0	110.0	132.0
1991	54.0	55.0	66.0	65.0	60.0	65.0	96.000000	55.000000	71.0	63.0	74.0	106.0
1992	34.0	47.0	56.0	53.0	53.0	55.0	67.000000	52.000000	46.0	51.0	58.0	91.0
1993	33.0	40.0	46.0	45.0	41.0	55.0	57.000000	54.000000	46.0	52.0	48.0	77.0
1994	30.0	35.0	42.0	48.0	44.0	45.0	45.333333	45.666667	46.0	51.0	63.0	84.0
1995	30.0	39.0	45.0	52.0	28.0	40.0	62.000000	NaN	NaN	NaN	NaN	NaN





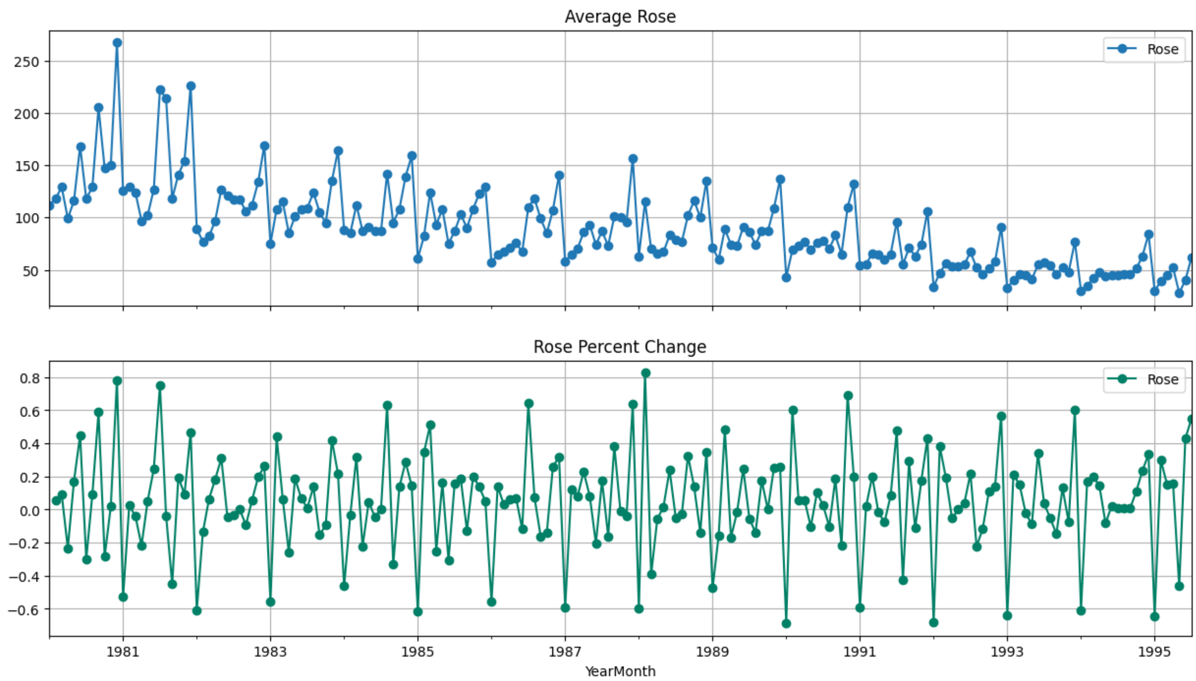
- Plot the Empirical Cumulative Distribution.



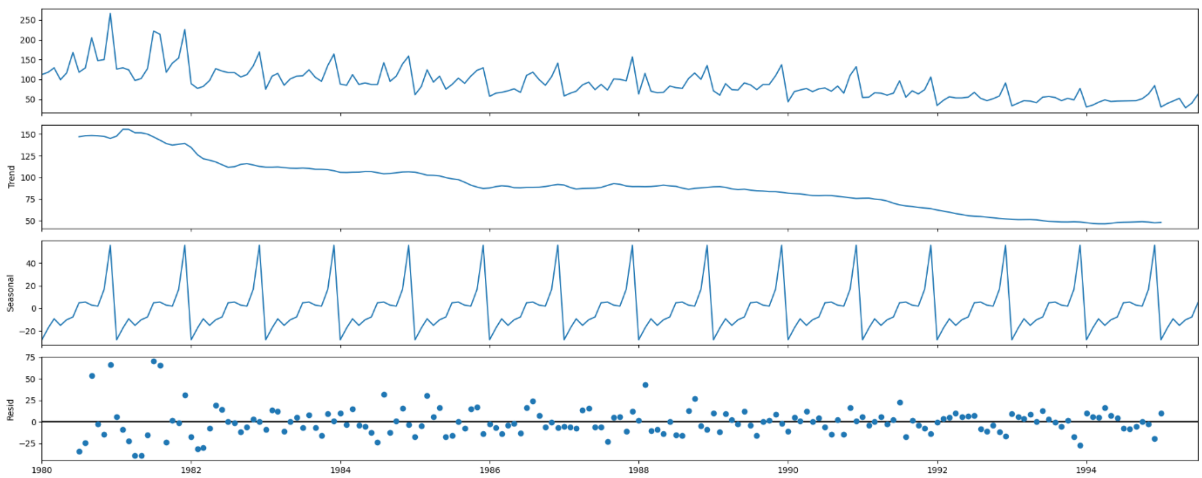
The interference from Empirical Cumulative Distribution is:

- Highest value is 250.
- Approximately 80% of sales is less than 150.

- **Average Rose Sales and Average Rose percentage**

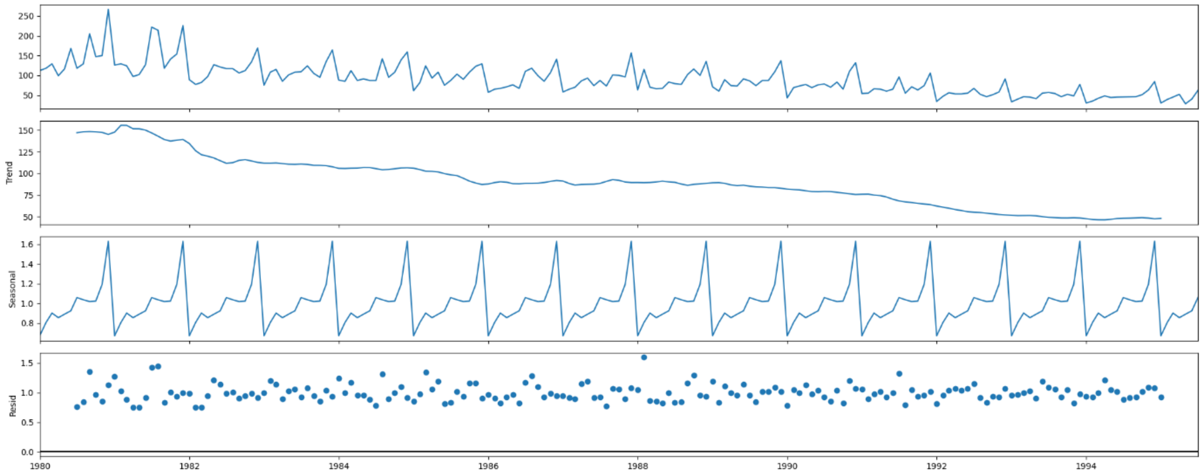


- **Decomposition --- Additive:**



- Decomposition by additive plot shows that:
  - It shows that the trend has been declined after 1981.
  - Peak year was 1981
  - Residue is spread like showing a pattern
  - Seasonality and trend is present.

- **Decomposition --- Multiplicative:**



- Decomposition by multiplicative plot shows that:
  - It shows that the trend has been declined after 1981.
  - Peak year was 1981
  - Residue is between 0 and 1 but in additive it is between 0 and 50 that is high.
  - Seasonality and trend is present.
  - Multiplicative decomposition is better in this case as the residue is between the range 0 and 1.

- **Trend,Seasonality and Residue**

```
Residual
 YearMonth
1980-01-01          NaN
1980-02-01          NaN
1980-03-01          NaN
1980-04-01          NaN
1980-05-01          NaN
1980-06-01          NaN
1980-07-01   -33.980241
1980-08-01   -24.624686
1980-09-01    53.850314
1980-10-01    -2.955241
1980-11-01   -14.263575
1980-12-01    66.161425
Name: resid, dtype: float64
```

```
Trend
 YearMonth
1980-01-01           NaN
1980-02-01           NaN
1980-03-01           NaN
1980-04-01           NaN
1980-05-01           NaN
1980-06-01           NaN
1980-07-01    147.083333
1980-08-01    148.125000
1980-09-01    148.375000
1980-10-01    148.083333
1980-11-01    147.416667
1980-12-01    145.125000
Name: trend, dtype: float64

Seasonality
 YearMonth
1980-01-01   -27.908647
1980-02-01   -17.435632
1980-03-01    -9.285830
1980-04-01   -15.098330
1980-05-01   -10.196544
1980-06-01    -7.678687
1980-07-01     4.896908
1980-08-01     5.499686
1980-09-01     2.774686
1980-10-01     1.871908
1980-11-01    16.846908
1980-12-01    55.713575
Name: seasonal, dtype: float64
```

3. Split the data into training and test. The test data should start in 1991.

The dataset is being split into training and test dataset. The test data set starts from 1991.

The train dataset has 132 rows and 1 column.

The test dataset has 55 rows and 1 column.

- **Dataset of train as shown below**:

Training Data

YearMonth	Rose
1980-01-01	112.0
1980-02-01	118.0
1980-03-01	129.0
1980-04-01	99.0
1980-05-01	116.0
...	...
1990-08-01	70.0
1990-09-01	83.0
1990-10-01	65.0
1990-11-01	110.0
1990-12-01	132.0

132 rows × 1 columns

Test dataset:

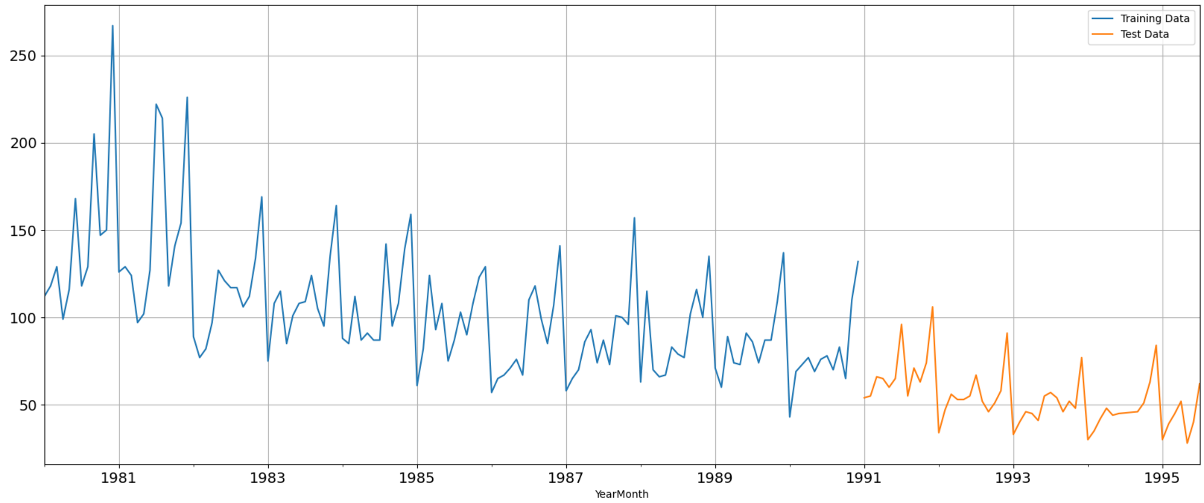
Test Data

YearMonth	Rose
1991-01-01	54.000000
1991-02-01	55.000000
1991-03-01	66.000000
1991-04-01	65.000000
1991-05-01	60.000000
1991-06-01	65.000000
1991-07-01	96.000000
1991-08-01	55.000000
1991-09-01	71.000000
1991-10-01	63.000000
1991-11-01	74.000000
1991-12-01	106.000000
1992-01-01	34.000000
1992-02-01	47.000000
1992-03-01	56.000000

Train datasets describe:

	Rose
count	132.000000
mean	104.939394
std	36.171508
min	43.000000
25%	77.750000
50%	99.500000
75%	121.500000
max	267.000000

Plot for training and test dataset:



4. Build all the exponential smoothing models on the training data and evaluate the model using RMSE on the test data. Other models such as regression, naive forecast models and simple average models. should also be built on the training data and check the performance on the test data using RMSE.

The below models are built on Training and test dataset.

- Linear Regression
- Naive Approach
- Simple Average
- Moving Average (MA)
- Simple Exponential Smoothing
- Double Exponential Smoothing (Holt's Model)
- Triple Exponential Smoothing (Holt - Winter's Model)

LINEAR REGRESSION:

- Few rows of training and test dataset:

First few rows of Training Data

YearMonth	Rose	time
1980-01-01	112.0	1
1980-02-01	118.0	2
1980-03-01	129.0	3
1980-04-01	99.0	4
1980-05-01	116.0	5

Last few rows of Training Data

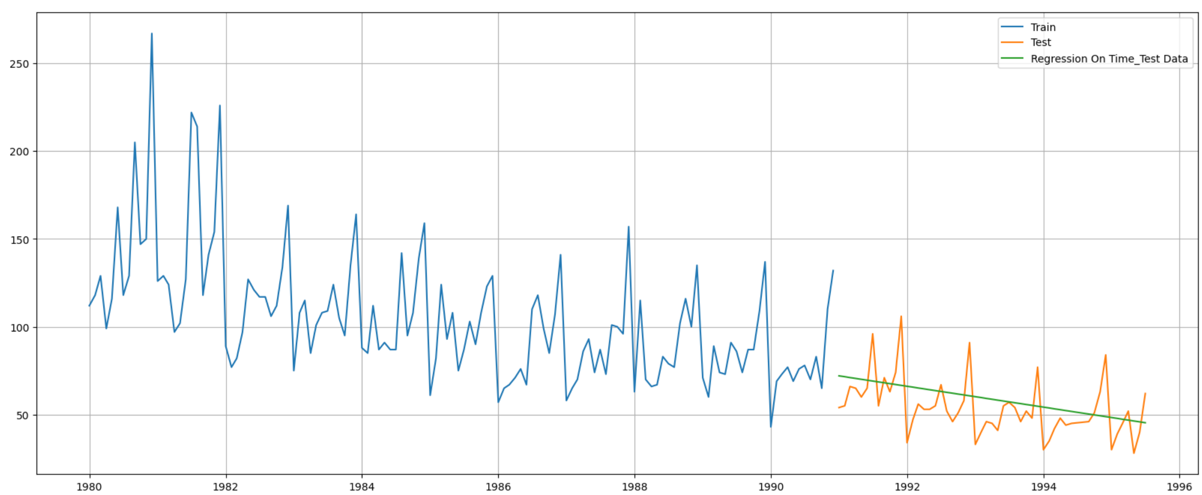
YearMonth	Rose	time
1990-08-01	70.0	128
1990-09-01	83.0	129
1990-10-01	65.0	130
1990-11-01	110.0	131
1990-12-01	132.0	132

First few rows of Test Data

YearMonth	Rose	time
1991-01-01	54.0	133
1991-02-01	55.0	134
1991-03-01	66.0	135
1991-04-01	65.0	136
1991-05-01	60.0	137

Last few rows of Test Data

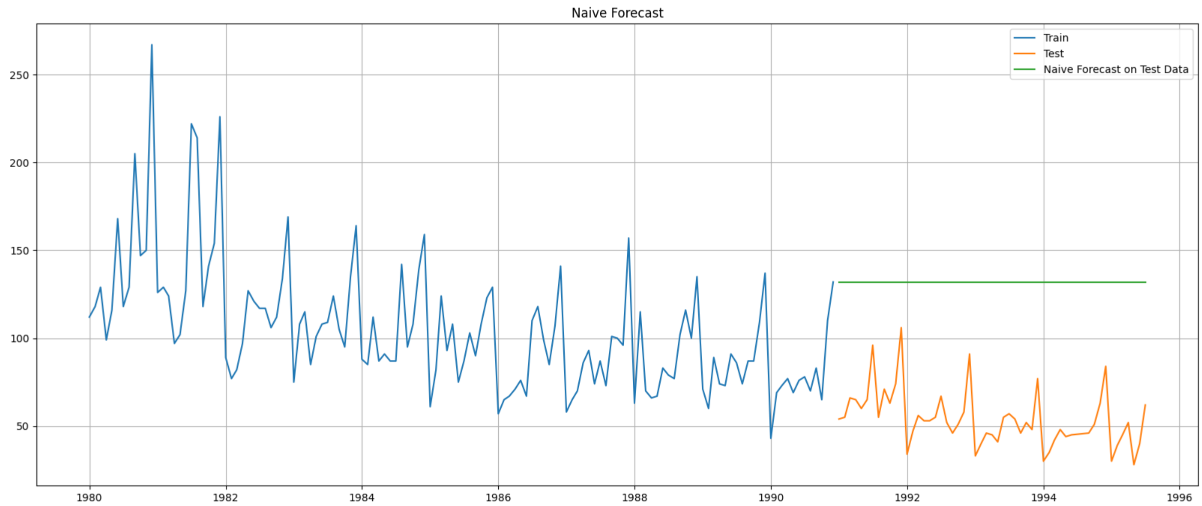
YearMonth	Rose	time
1995-03-01	45.0	183
1995-04-01	52.0	184
1995-05-01	28.0	185
1995-06-01	40.0	186
1995-07-01	62.0	187



- The green line is the prediction made by Linear Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 15.26.

	Test RMSE
Linear Regression	15.268955

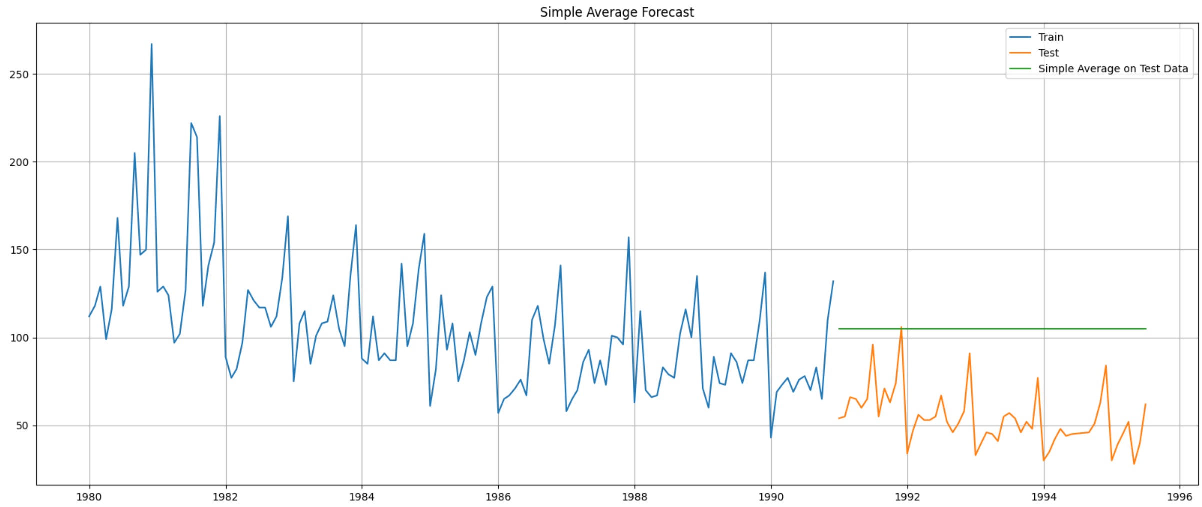
NAIVE'S MODEL:



- The green line is the prediction made by Naive's Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 79.71. The less the RMSE the better the model.

NaiveModel	79.718773

SIMPLE AVERAGE FORECAST MODEL:

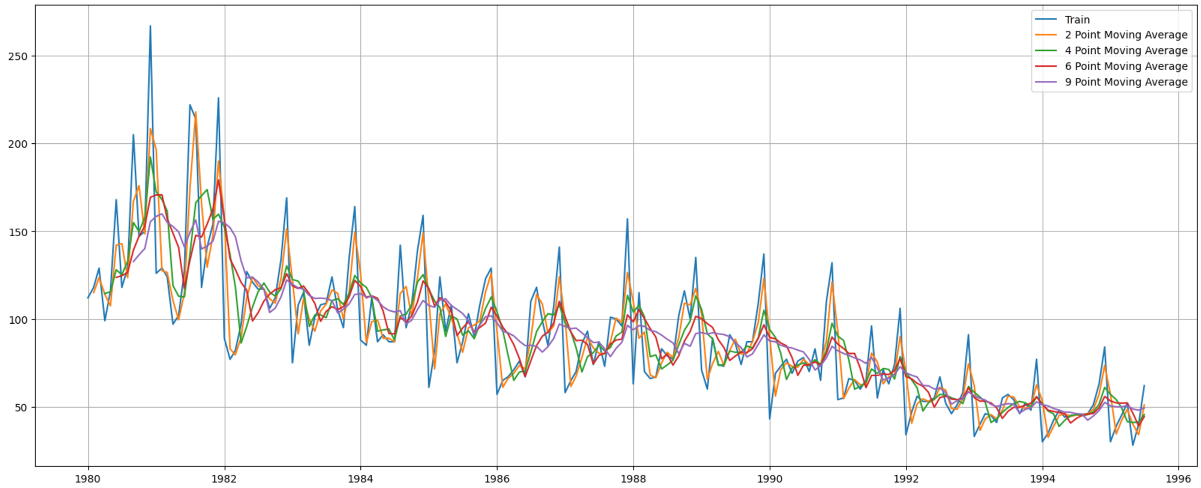


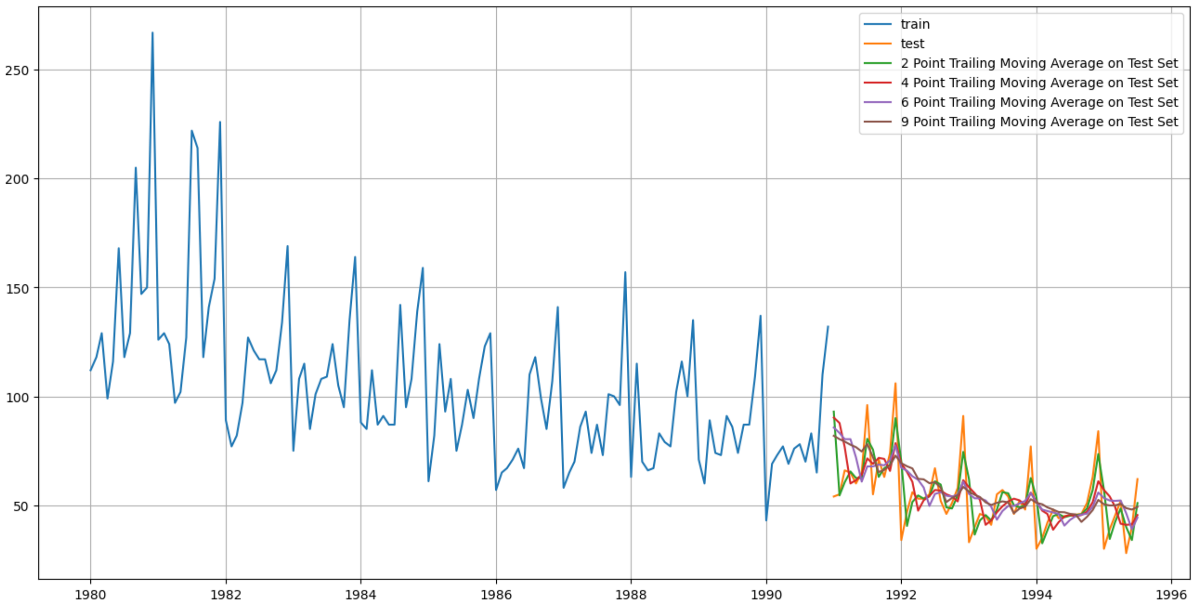
- The green line is the prediction made by Simple Average Forecast Model as we can see that the prediction made by the linear model is not good as shown in graph also as the predicted values is far away from the actual values.
- RMSE values calculated for the model is 53.46. The less the RMSE the better the model.

SimpleAverageModel	53.460570

- ## MOVING AVERAGE:

	Rose	Trailing_2	Trailing_4	Trailing_6	Trailing_9
YearMonth					
1980-01-01	112.0	NaN	NaN	NaN	NaN
1980-02-01	118.0	115.0	NaN	NaN	NaN
1980-03-01	129.0	123.5	NaN	NaN	NaN
1980-04-01	99.0	114.0	114.5	NaN	NaN
1980-05-01	116.0	107.5	115.5	NaN	NaN





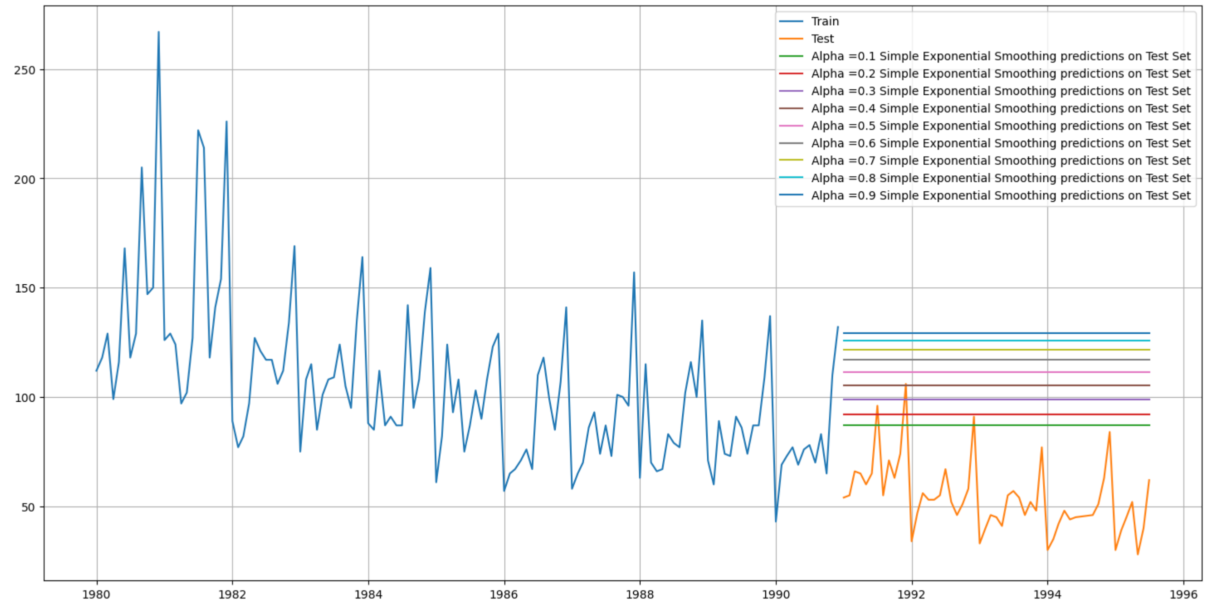
A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values. Higher the rolling window, smoother will be its curve more values are being taken into account.

- RMSE values calculated for the model are as below. The less the RMSE the better the model.

2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630

- ## SIMPLE EXPONENTIAL:

Taken all values from 0.1 to 0.9 to find the best alpha value for SIMPLE EXPONENTIAL which has less RMSE .



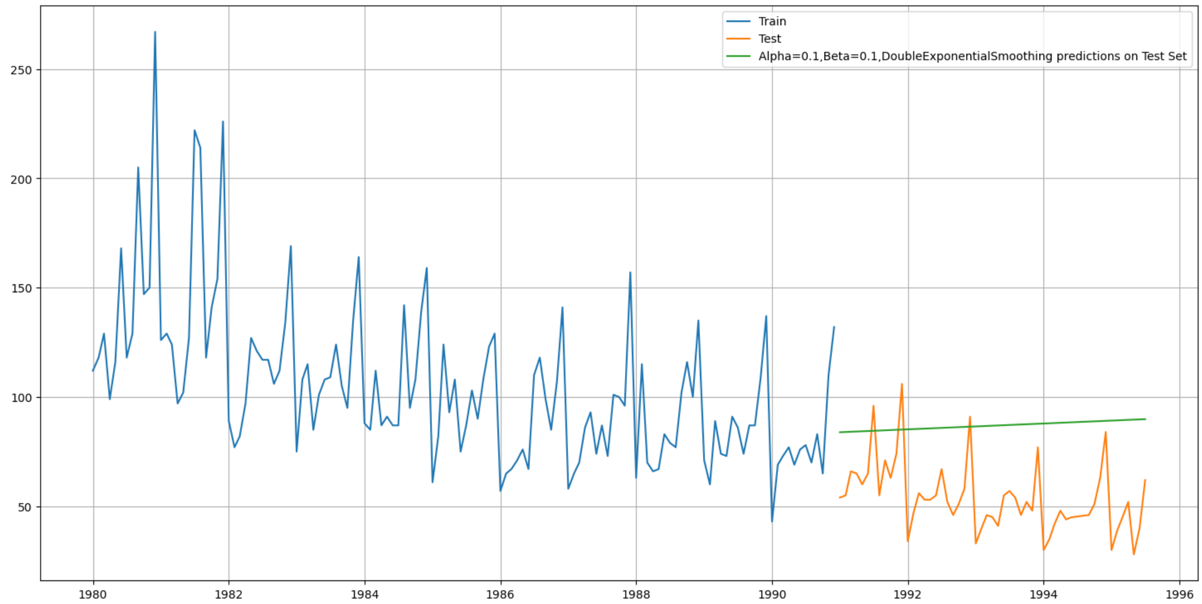
RMSE:

The alpha value 0.1 is giving us less RMSE that is 36.82 in all the apha values .

	Alpha Values	Train RMSE	Test RMSE
0	0.1	31.815610	36.828033
1	0.2	31.979391	41.361876
2	0.3	32.470164	47.504821
3	0.4	33.035130	53.767406
4	0.5	33.682839	59.641786
5	0.6	34.441171	64.971288
6	0.7	35.323261	69.698162
7	0.8	36.334596	73.773992
8	0.9	37.482782	77.139276

- **Double Exponential Smoothing (Holt's Model) :**

Taken all values from 0.1 to 0.9 to find the best alpha , beta value for DOUBLE EXPONENTIAL which has less RMSE .



RMSE:

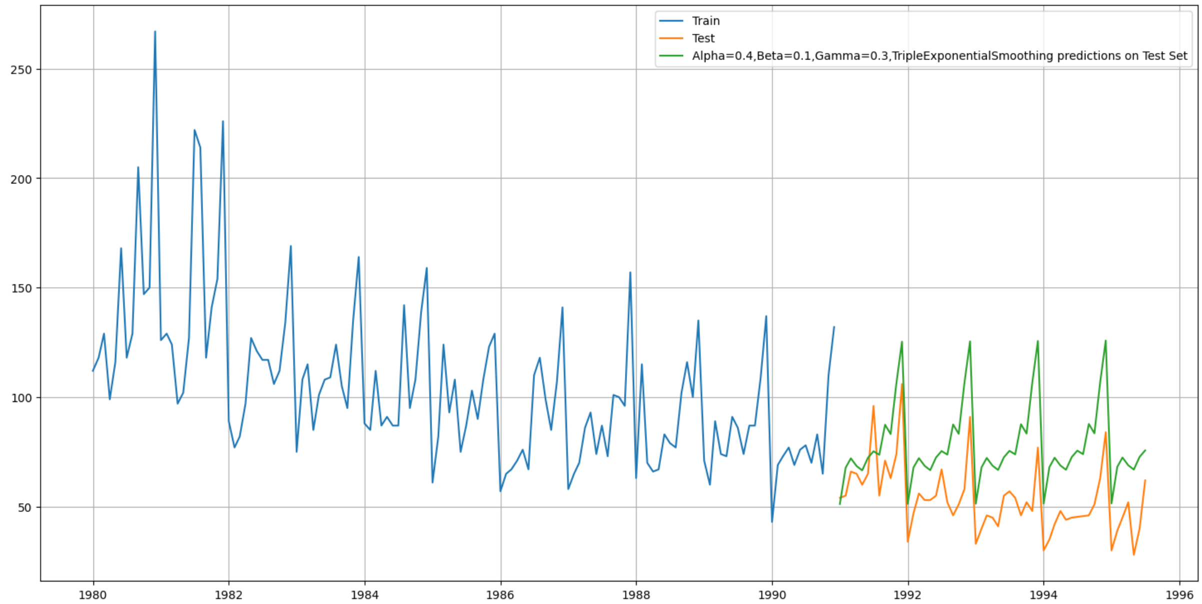
The alpha value and beta 0.1 is giving us less RMSE that is 36.92 in all the apha , beta values . So best value for alpha, beta is 0.1.

	Alpha Values	Beta Values	Train RMSE	Test RMSE
0	0.1	0.1	34.439111	36.923416
1	0.1	0.2	33.450729	48.688648
2	0.1	0.3	33.145789	78.156641
3	0.1	0.4	33.262191	99.583473
4	0.1	0.5	33.688415	124.269726
...	...	...	...	...
95	1.0	0.6	51.831610	801.680218
96	1.0	0.7	54.497039	841.892573
97	1.0	0.8	57.365879	853.965537
98	1.0	0.9	60.474309	834.710935
99	1.0	1.0	63.873454	780.079579

100 rows × 4 columns

- **Triple Exponential Smoothing (Holt Winter's Model) :**

Taken all values from 0.1 to 0.9 to find the best alpha , beta , gamma value for triple exponential smoothing to see which has less RMSE . We can see that the predicted value that is green is fitting the actual values much better than the other models



- **RMSE:**

The alpha value 0.2, beta 0.7 and gamma 0.3 is giving us less RMSE that is 8.70 in all the apha , beta and gamma values . So best value for alpha, beta , gamma is that only.

	Alpha Values	Beta Values	Gamma Values	Train RMSE	Test RMSE	Method
2136	0.2	0.7	0.2	24.042290	8.702460	tm_sm
1010	0.1	0.2	0.1	19.770392	9.223504	ta_sm
1011	0.1	0.2	0.2	20.253487	9.496152	ta_sm
1151	0.2	0.6	0.2	23.129850	9.565988	ta_sm
1012	0.1	0.2	0.3	20.871304	9.888106	ta_sm

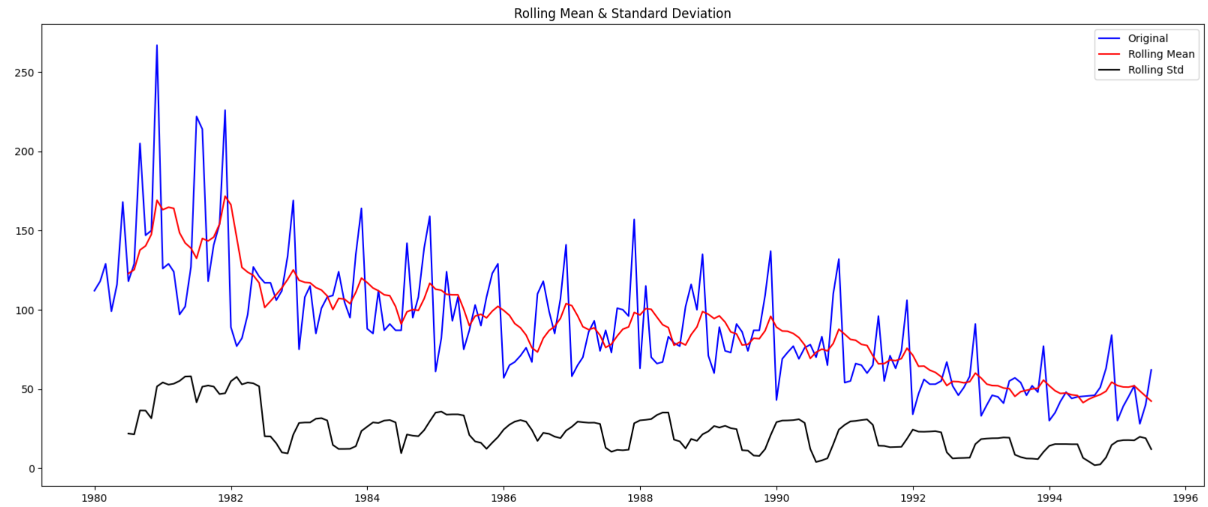
5. Check for the stationarity of the data on which the model is being built on using appropriate statistical tests and mention the hypothesis for the statistical test. If the data is found to be non-stationary, take appropriate steps to make it stationary. Check the new data for stationarity and comment. Note: Stationarity should be checked at alpha = 0.05.

For checking the series is stationary or not we have to use Augumented Dickey – Fuller test for the same.

The hypothesis for this is :

If the value is less than 0.05 then the series is stationary and good to move for further ARIMA/SARIMA Model.

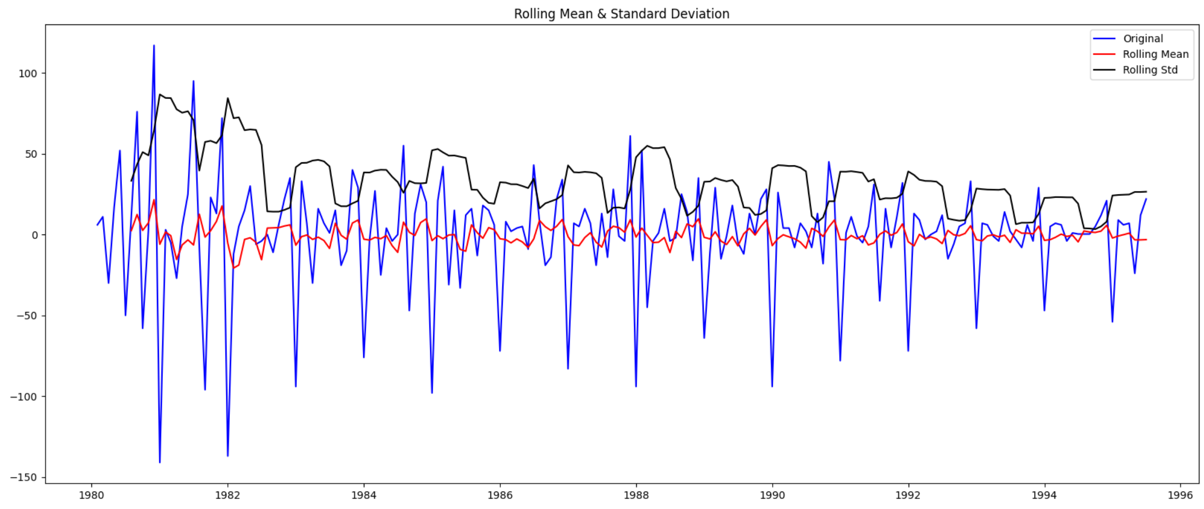
If the value of p value is more than 0.05 then we fail to reject the null hypothesis and the series is not stationary and can't proceed with ARIMA/SARIMA model.




```
Results of Dickey-Fuller Test:
Test Statistic                  -1.876699
p-value                          0.343101
#Lags Used                      13.000000
Number of Observations Used    173.000000
Critical Value (1%)             -3.468726
Critical Value (5%)             -2.878396
Critical Value (10%)            -2.575756
dtype: float64
```

Let us take a difference of order 1 and check whether the Time Series is stationary or not.

We used .diff() function on the existing series without any argument, implying the default diff value of 1 and also dropped the NaN values, since differencing of order 1 would generate the first value as NaN which need to be dropped



We can see that now the p value is 1.810895e-12 that is much smaller than the 0.05 so we fail to reject the null hypothesis and considering the series as stationary and good to move further for ARIMA / SARIMA Model as the series is stationary.

```
Results of Dickey-Fuller Test:
Test Statistic                -8.044392e+00
p-value                        1.810895e-12
#Lags Used                     1.200000e+01
Number of Observations Used    1.730000e+02
Critical Value (1%)           -3.468726e+00
Critical Value (5%)           -2.878396e+00
Critical Value (10%)          -2.575756e+00
dtype: float64
```

6. Build an automated version of the ARIMA/SARIMA model in which the parameters are selected using the lowest Akaike Information Criteria (AIC) on the training data and evaluate this model on the test data using RMSE.

The values of p,q,d where p is the order of AR , q is the order of Moving average and d is the difference that will make the series stationary for this a for loop has been there .

```
Some parameter combinations for the Model...
Model: (0, 1, 1)
Model: (0, 1, 2)
Model: (0, 1, 3)
Model: (1, 1, 0)
Model: (1, 1, 1)
Model: (1, 1, 2)
Model: (1, 1, 3)
Model: (2, 1, 0)
Model: (2, 1, 1)
Model: (2, 1, 2)
Model: (2, 1, 3)
Model: (3, 1, 0)
Model: (3, 1, 1)
Model: (3, 1, 2)
Model: (3, 1, 3)
```

Less the AIC we will take that model in this case 2,1,3 has the lowest AIC so we need to sort the AIC.

```
ARIMA(0, 1, 0) - AIC:1333.1546729124348
ARIMA(0, 1, 1) - AIC:1282.309831974832
ARIMA(0, 1, 2) - AIC:1279.6715288535784
ARIMA(0, 1, 3) - AIC:1280.545376173466
ARIMA(1, 1, 0) - AIC:1317.350310538146
ARIMA(1, 1, 1) - AIC:1280.5742295380046
ARIMA(1, 1, 2) - AIC:1279.8707234231929
ARIMA(1, 1, 3) - AIC:1281.8707223309984
ARIMA(2, 1, 0) - AIC:1298.6110341604945
ARIMA(2, 1, 1) - AIC:1281.5078621868563
ARIMA(2, 1, 2) - AIC:1281.870722226456
ARIMA(2, 1, 3) - AIC:1274.6951271827177
ARIMA(3, 1, 0) - AIC:1297.481091727167
ARIMA(3, 1, 1) - AIC:1282.4192776271927
ARIMA(3, 1, 2) - AIC:1283.7207405977094
ARIMA(3, 1, 3) - AIC:1278.6580044819445
```

After the sort we found that Less the AIC we will take that model in this case 2,1,3 has the lowest AIC.

	param	AIC
11	(2, 1, 3)	1274.695127
15	(3, 1, 3)	1278.658004
2	(0, 1, 2)	1279.671529
6	(1, 1, 2)	1279.870723
3	(0, 1, 3)	1280.545376
5	(1, 1, 1)	1280.574230
9	(2, 1, 1)	1281.507862
10	(2, 1, 2)	1281.870722
7	(1, 1, 3)	1281.870722
1	(0, 1, 1)	1282.309832
13	(3, 1, 1)	1282.419278
14	(3, 1, 2)	1283.720741
12	(3, 1, 0)	1297.481092
8	(2, 1, 0)	1298.611034
4	(1, 1, 0)	1317.350311
0	(0, 1, 0)	1333.154673

The summary report for the ARIMA Model with values (2,1,3) as p,q,d respectively.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                   Rose   No. Observations:                  132
Model:                 ARIMA(2, 1, 3)   Log Likelihood                -631.348
Date:                Sat, 24 Feb 2024   AIC                           1274.695
Time:                        18:43:34   BIC                           1291.946
Sample:                    01-01-1980   HQIC                          1281.705
                         - 12-01-1990
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -1.6779      0.084    -20.034      0.000      -1.842      -1.514
ar.L2         -0.7288      0.084     -8.702      0.000      -0.893      -0.565
ma.L1          1.0447      0.644      1.622      0.105      -0.217       2.307
ma.L2         -0.7718      0.134     -5.775      0.000      -1.034      -0.510
ma.L3         -0.9046      0.584     -1.549      0.121      -2.049       0.240
sigma2       858.8436    541.924      1.585      0.113    -203.308    1920.995
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):                24.45
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               0.40   Skew:                             0.71
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.57
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

- **RMSE for AUTO_ARIMA:**

Auto_ARIMA	36.815186

- ## SARIMA MODEL

SARIMA utilizes a variety of auto-regression (AR) and moving average (MA) models, as well as differencing, to capture trends and seasonality in data.

```
Examples of some parameter combinations for Model...
Model: (0, 1, 1)(0, 0, 1, 12)
Model: (0, 1, 2)(0, 0, 2, 12)
Model: (0, 1, 3)(0, 0, 3, 12)
Model: (1, 1, 0)(1, 0, 0, 12)
Model: (1, 1, 1)(1, 0, 1, 12)
Model: (1, 1, 2)(1, 0, 2, 12)
Model: (1, 1, 3)(1, 0, 3, 12)
Model: (2, 1, 0)(2, 0, 0, 12)
Model: (2, 1, 1)(2, 0, 1, 12)
Model: (2, 1, 2)(2, 0, 2, 12)
Model: (2, 1, 3)(2, 0, 3, 12)
Model: (3, 1, 0)(3, 0, 0, 12)
Model: (3, 1, 1)(3, 0, 1, 12)
Model: (3, 1, 2)(3, 0, 2, 12)
Model: (3, 1, 3)(3, 0, 3, 12)
```

```
SARIMA(0, 1, 0)x(0, 0, 0, 12) - AIC:1323.9657875279158
SARIMA(0, 1, 0)x(0, 0, 1, 12) - AIC:1145.4230827207298
SARIMA(0, 1, 0)x(0, 0, 2, 12) - AIC:976.4375296380895
SARIMA(0, 1, 0)x(0, 0, 3, 12) - AIC:3537.579168905919
SARIMA(0, 1, 0)x(1, 0, 0, 12) - AIC:1139.921738995602
SARIMA(0, 1, 0)x(1, 0, 1, 12) - AIC:1116.0207869386172
SARIMA(0, 1, 0)x(1, 0, 2, 12) - AIC:969.691363575225
SARIMA(0, 1, 0)x(1, 0, 3, 12) - AIC:4554.32909051064
SARIMA(0, 1, 0)x(2, 0, 0, 12) - AIC:960.8812220353041
SARIMA(0, 1, 0)x(2, 0, 1, 12) - AIC:962.8794540697556
SARIMA(0, 1, 0)x(2, 0, 2, 12) - AIC:955.5735408945757
SARIMA(0, 1, 0)x(2, 0, 3, 12) - AIC:4397.822817992214
SARIMA(0, 1, 0)x(3, 0, 0, 12) - AIC:850.7535403931095
SARIMA(0, 1, 0)x(3, 0, 1, 12) - AIC:851.7482702748039
SARIMA(0, 1, 0)x(3, 0, 2, 12) - AIC:850.53041361288
SARIMA(0, 1, 0)x(3, 0, 3, 12) - AIC:3467.855628476979
SARIMA(0, 1, 1)x(0, 0, 0, 12) - AIC:1263.5369097383966
SARIMA(0, 1, 1)x(0, 0, 1, 12) - AIC:1098.5554825918337
SARIMA(0, 1, 1)x(0, 0, 2, 12) - AIC:923.631404938385
SARIMA(0, 1, 1)x(0, 0, 3, 12) - AIC:3915.4769311640416
SARIMA(0, 1, 1)x(1, 0, 0, 12) - AIC:1095.793632491823
SARIMA(0, 1, 1)x(1, 0, 1, 12) - AIC:1054.7434330946953
SARIMA(0, 1, 1)x(1, 0, 2, 12) - AIC:918.8573483297299
SARIMA(0, 1, 1)x(1, 0, 3, 12) - AIC:3917.4099549077478
SARIMA(0, 1, 1)x(2, 0, 0, 12) - AIC:914.5982866535833
SARIMA(0, 1, 1)x(2, 0, 1, 12) - AIC:915.333243046168
SARIMA(0, 1, 1)x(2, 0, 2, 12) - AIC:901.1988272651953
SARIMA(0, 1, 1)x(2, 0, 3, 12) - AIC:3887.5888930228098
SARIMA(0, 1, 1)x(3, 0, 0, 12) - AIC:798.588976481104
SARIMA(0, 1, 1)x(3, 0, 1, 12) - AIC:800.4844931540345
SARIMA(0, 1, 1)x(3, 0, 2, 12) - AIC:801.0595269469408
```

After the sort we found that Less the AIC we will take that model in this case (3,1,1,) (3,0,2,12) has the lowest AIC.

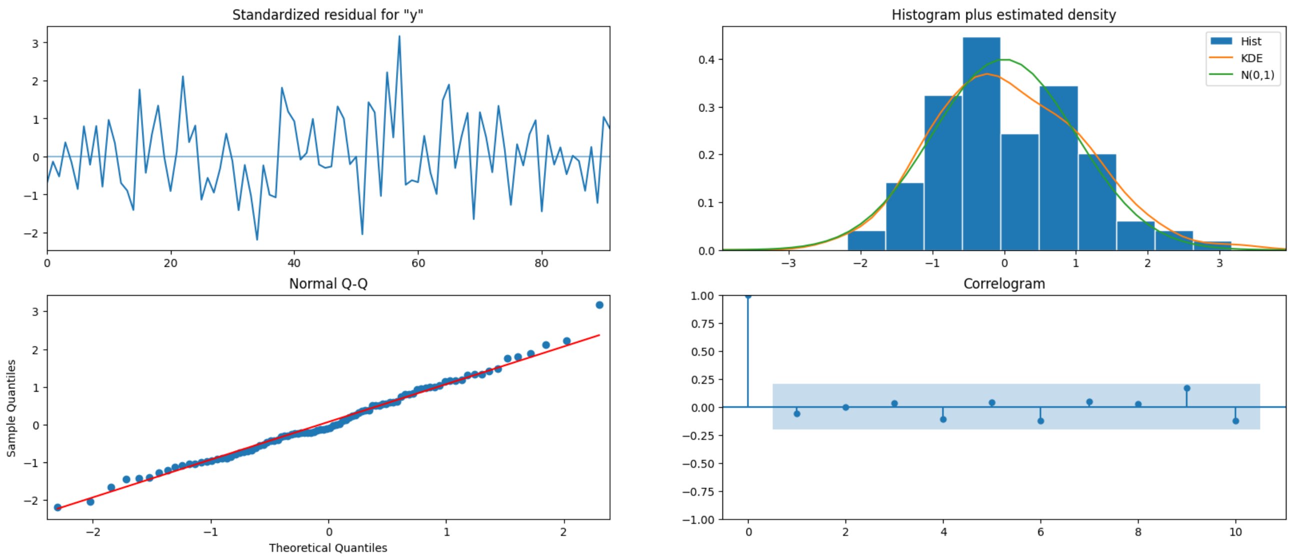
	param	seasonal	AIC
222	(3, 1, 1)	(3, 0, 2, 12)	774.400287
238	(3, 1, 2)	(3, 0, 2, 12)	774.880936
220	(3, 1, 1)	(3, 0, 0, 12)	775.426699
221	(3, 1, 1)	(3, 0, 1, 12)	775.495330
252	(3, 1, 3)	(3, 0, 0, 12)	775.561018

The summary report for the ARIMA Model with values (3,1,1,) (3,0,2,12) model.

```
                                     SARIMAX Results
==============================================================================================
Dep. Variable:                                   y   No. Observations:                     132
Model:             SARIMAX(3, 1, 1)x(3, 0, [1, 2], 12)   Log Likelihood                  -377.200
Date:                             Sat, 24 Feb 2024   AIC                                774.400
Time:                                     18:53:19   BIC                                799.618
Sample:                                          0   HQIC                               784.578
                                             - 132
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0464      0.126      0.367      0.714      -0.201       0.294
ar.L2         -0.0060      0.120     -0.050      0.960      -0.241       0.229
ar.L3         -0.1808      0.098     -1.838      0.066      -0.374       0.012
ma.L1         -0.9370      0.067    -13.903      0.000      -1.069      -0.805
ar.S.L12       0.7639      0.165      4.640      0.000       0.441       1.087
ar.S.L24       0.0840      0.159      0.527      0.598      -0.229       0.397
ar.S.L36       0.0727      0.095      0.764      0.445      -0.114       0.259
ma.S.L12      -0.4969      0.250     -1.988      0.047      -0.987      -0.007
ma.S.L24      -0.2191      0.210     -1.044      0.296      -0.630       0.192
sigma2       192.1390     39.627      4.849      0.000     114.471     269.807
===================================================================================
Ljung-Box (L1) (Q):                   0.30   Jarque-Bera (JB):                 1.64
Prob(Q):                              0.58   Prob(JB):                         0.44
Heteroskedasticity (H):               1.11   Skew:                             0.33
Prob(H) (two-sided):                  0.77   Kurtosis:                         3.03
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Graphs for the residual to determine if any further information can be extracted or all the usable information has already been extracted .



y	mean	mean_se	mean_ci_lower	mean_ci_upper
0	55.237188	13.907058	27.979855	82.494520
1	68.122541	13.990531	40.701604	95.543478
2	67.909380	14.011597	40.447154	95.371605
3	66.786145	14.098878	39.152852	94.419438
4	69.761986	14.108245	42.110334	97.413638

- **RMSE for SARIMA:**

(3,1,1),(3,0,2,12),Auto_SARIMA	18.882146

7. Build a table (create a data frame) with all the models built along with their corresponding parameters and the respective RMSE values on the test data.

As we can see that the model that has the lowest RMSE is Exponential Smoothing with 0.2 as Alpha , 0.7 as beta and 0.2 as Gamma with 8.70 is the best.

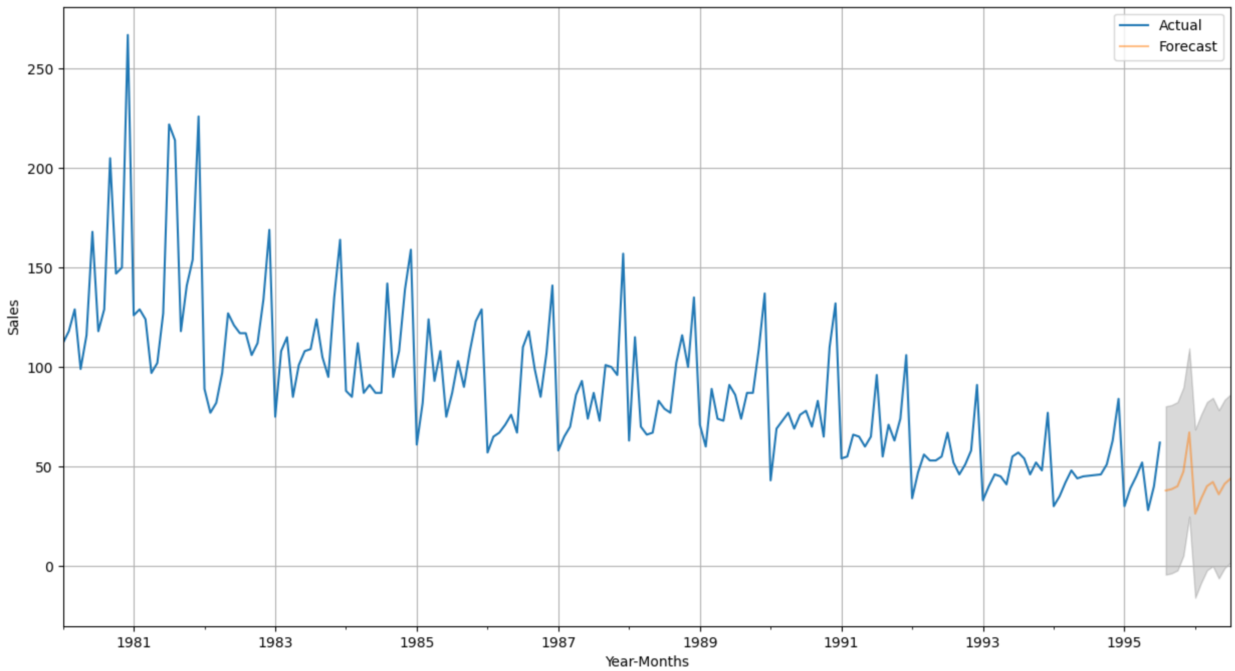
	Test RMSE
Alpha=0.2,Beta=0.7,Gamma=0.2,TripleExponentialSmoothing	8.702460
2pointTrailingMovingAverage	11.529278
4pointTrailingMovingAverage	14.451403
6pointTrailingMovingAverage	14.566327
9pointTrailingMovingAverage	14.727630
(2,1,2)(2,1,2,12),Manual_SARIMA	15.168791
Linear Regression	15.268955
(3,1,1),(3,0,2,12),Auto_SARIMA	18.882146
Auto_ARIMA	36.815186
Alpha=0.1,SimpleExponentialSmoothing	36.828033
ARIMA(3,1,3)	36.871197
Alpha Value = 0.1, beta value = 0.1, DoubleExponentialSmoothing	36.923416
Alpha=0.08621,Beta=1.3722,Gamma=0.4763,TrippleExponentialSmoothing_Auto_Fit	37.592212
SimpleAverageModel	53.460570
NaiveModel	79.718773

8. Based on the model-building exercise, build the most optimum model(s) on the complete data and predict 12 months into the future with appropriate confidence intervals/bands.

We can see that the optimum model with lowest RMSE is exponential smoothing so this model will be ideal for making predictions. Considering Exponential smoothing model ideal we will make prediction as below:

	Rose_Sale_Prediction
1995-08-01	37.915551
1995-09-01	38.575089
1995-10-01	40.066433
1995-11-01	47.417515
1995-12-01	67.106532
1996-01-01	26.260601
1996-02-01	33.743810
1996-03-01	40.102270
1996-04-01	42.188188
1996-05-01	36.003167
1996-06-01	41.266405
1996-07-01	43.913350

The Sales prediction of Rose wine graph with confidence level as shown below:



9. Comment on the model thus built and report your findings and suggest the measures that the company should be taking for future sales.

- There is peak in 1980-1981. Though outliers are also present in mostly all years in Yearly.
- From the monthly we can inference that the sales of rose wine is mostly high in December and lowest in January.
- Outliers are present in June, July, August, September and December.
- We can see that the month January has the lowest sale and December has the highest sale.
- Some bumper offers should be launched during the month of April to June to increase the sale at that time.
- May to June there is average sale not high not low.
- The year 1981 was the year with the highest number of sales.
- This trend is expected to continue in the future as well, based on the prediction with most optimal model.